Towards the Study of Human Emotions Through Social Media Contents

Manuel Barreto Lima Reis

Instituto Superior Técnico

Abstract. This work proposes on building a automated method based on neural embeddings and regression models for inferring the emotional content, in terms of continuous scores of valence, arousal, and dominance for textual contents (i.e., words in English, Portuguese, Spanish, Italian, German and short texts in English). Later on, leveraging on the aforementioned methods and social media contents from Twitter, are proposed two applications: (1) assessing well-being levels for populations across continental USA (excluding Alaska), and (2) studying the relationship between the emotional content of a given message in Twitter and how this message spreads in the referred social network, e.g., trying to see whether markedly negative messages reach a larger number of users than neutral messages.

Keywords: Natural Language Processing; Sentiment Analysis; Information Diffusion; Social Media

1 Introduction

Emotions play a major role in human interactions, and are therefore, intrinsically embedded in either textual or oral communication between human beings. The emotional characteristics of a message, contribute consequently, to its semantics. Besides, social media, is nowadays, a topic of large interest mainly due to the fact that is an easy and cheap way of analyzing human behaviour and interaction. For instance, several studies have been studying the relation between contents in social media, and real measurable metrics, such as the happiness of populations [Loff et al., 2015] or the weather [Li et al., 2014]. User-generated contents from social networks, are produced every day and around the globe on an unprecedented scale. Having such a large set of contents, we can leverage on these to ask a myriad of questions. Another topic of great interest, has been leveraging on these contents and studying how they diffuse from user to user, and around the world.

The focus of the work presented in this paper was developing a method for extraction of emotional scores (in a continuous scale) from words (i.e., in English, Spanish, Portuguese, Italian and German), and textual messages, in terms of Valence, Arousal and Dominance. Then relying on this method and on a dataset of tweets, were produced two application: (1) Inferring well-being of populations across USA, and (2) studying the relation between these scores, and the way the messages associated to the aforementioned scores spread in Twitter, according to some diffusion metrics (e.g., geographic coverage of a message).

The rest of the paper is organized as follows: Section 2 describes some previous and related work. Section 3 presents a method which predicts emotional scores of unseen words in English and other languages (i.e., Spanish, Portuguese, Italian and German), as well of short texts. Then, Section 4 leverages on the aforementioned method to predict emotional ratings for tweets and: (1) tries to assess well-being of populations across continental USA (excluding Alaska), and (2) studies the relationship between these emotional ratings and some metrics associated to the diffusion of the tweet that originated those predictions. Finally, Section 5 summarizes the mains aspects discussed in this paper, and points some possible future work directions.

2 Related Work

In recent years, researchers developed new studies for collecting human emotional ratings for large sets of words, e.g. through crowdsourcing methodologies [Årup Nielsen, 2011, Dodds et al., 2011, Warriner et al., 2013], or to investigate whether affective norms can be estimated via automatic procedures. Given a small seed lexicon with emotional ratings, a number of techniques have actually been proposed for extending it [Turney and Littman, 2003, Pitel and Grefenstette, 2008, Yu et al., 2013, Grefenstette et al., 2006].

For instance, Bestgen and Vincze [2012] used Latent Semantic Analysis (LSA), a well-known algorithm for quantifying the degree to which terms are associated in a large text corpus, in order to automatically produce estimates of the valence, arousal and dominance of words.

Recchiaa and Louwerse [2014] attempted to further improve results, by integrating into the prediction models additional variables that in the literature have been shown to correlate with valence, arousal or dominance, and by using a more scalable approach based on Point-wise Mutual Information (PMI). The authors measured correlations of 0.74, 0.57 and 0.62 to ANEW valence, arousal and dominance ratings.

Manderaa et al. [2015] also researched the usage of machine learning techniques (i.e., k nearest neighbors and random forest regression) to extrapolate existing ratings to previously unrated words, levarage word representation built through (i) latent semantic analysis, (ii) Latent Dirichlet Allocation [Blei et al., 2003], (iii) PMI, and word2vec's skip-ngram model embeddings. The authors reported correlations of 0.478, 0.595 and 0.694 in 10-fold cross validation experiments with the ratings from Warriner et al. [2013], respectively in terms of valence, arousal and dominance.

Regarding the prediction of emotional scores for short texts, Paltoglou et al. [2013] have previously reported on a study in which subjects were asked to rate the emotional impact of 20 forum discussion posts, according to valence and arousal. These documents were later analyzed through the ANEW lexicon, estimating the emotional content of each document through the weighted geometric mean of the ANEW tokens found it the text. The authors report on correlations with the human assessments of 0.89 for valence and of 0.42 for arousal. Murphy [2014] also experimented with a simple procedure based on assigning documents to the average ANEW scores of their words, validating their results on data from the Affective Norms for English Texts (ANET) dataset Bradley and Lang [2007]. The authors report on correlations of 0.45 for valence, 0.31 for arousal, and 0.26 for dominance.

The methods presented in this paper differ significantly from the aforementioned approaches, by going beyond rule-based methods that use lexicons for text analysis, and by leveraging unsupervised embeddings for words, as well as, state-of-the-art regression models.

3 Predicting Affective Norms for Words and Short Texts

Human emotional ratings of valence, arousal, and dominance, for particular words, are nowadays frequently used within cognitive science, behavioral psychology and psycholinguistic research. These motional ratings are typically collected through interviews, by asking participants to rate words according to the emotional dimensions (i.e., valence, arousal and dominance) under consideration. A dimension of valence can be defined as the pleasantness of the stimulus, arousal can, in turn, be identified with the intensity of feeling being evoked by a particular word. Finally, a dimension of dominance can be identified with the degree to which the word makes the reader feel.

Collecting emotion norms from human raters is both expensive and time consuming. As a result, affective norms are available for only a few English words [Warriner et al., 2013, Bradley and Lang, 1999], are not available for proper nouns even in English [Recchiaa and Louwerse, 2014], and are sparse in other languages [Redondo et al., 2007, Soares et al., 2012, Montefinese et al., 2014, Schmidtke et al., 2014].

The aforementioned limitations have motivated researchers to seek automated procedures for estimating affective norms. In addition to this, in recent years, several unsupervised methods based on neural network architectures have been proposed to derive word embeddings from large corpora. In this context, word embeddings correspond to dense vector representations that implicitly capture syntactic and semantic properties of words (i.e., we have that a notion of semantic similarity, as well as other linguistic regularities, seem to be encoded in the embedding spaces resulting from these methods [Mikolov et al., 2013]).

Word representations based on neural network architectures have been shown to outperform other distributional similarity approaches [Baroni et al., 2014], and is a goal of this work to argue that these embedding vectors can be used as features to train a regression model for predicting the emotional properties for new words. Taking inspiration on recent developments within computational linguistics, was also explored the possibility of making cross-language extrapolations for psycholinguistic variables.

	ANEW					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
k-NN	0.869	0.944	0.673	0.876	0.731	**0.551**
Random Forest	0.787	1.303	0.518	1.014	0.663	0.643
Kernel Ridge	**0.908**	**0.715**	**0.738**	**0.804**	**0.757**	0.560

	Warriner et al.					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
k-NN	0.895	0.821	0.714	0.596	0.838	0.558
Random Forest	0.817	1.214	0.559	0.750	0.748	0.745
Kernel Ridge	**0.934**	**0.575**	**0.769**	**0.527**	**0.856**	**0.480**

Table 1: Obtained results when predicting ratings for words in the English ANEW lexicon Bradley and Lang [1999] and in the lexicon from Warriner et al. [2013]. The associated p-values for the Pearson product-moment correlation coefficient were always lower than 0.001.

3.1 Monolingual Results

The first experiments consisted in the usage of English words in the set of norms from Warriner et al. [2013] that did not appear in the ANEW corpus, as training data for predictive models that can later be used to estimate valence, arousal and dominance ratings for previously unseen words. Word embeddings were leveraged as features within different types of regression approaches. Later on, the same approach was followed by training the regression models with words from the ANEW corpus that did not appear in the lexicon from Warriner et al. [2013].

The aforementioned word representations were used together with three different types of forecasting models, namely a k nearest neighbor interpolation approach, random forest regression, and kernel ridge regression. The three approaches were implemented through the scikit-learn library [Pedregosa et al., 2011].

Table 1 presents the results obtained in our first set of experiments, both in terms of Pearson's correlation coefficient ρ, and in terms of the Mean Absolute Error (MAE). These correlation values are similar to those reported on the previous studies by Bestgen and Vincze [2012] and by Recchiaa and Louwerse [2014], even slightly superior. For comparison, correlations between the valence, arousal and dominance ratings given in the original ANEW, by Bradley and Lang [1999] and in the study by Warriner et al. [2013] are, respectively, of 0.953, 0.761 and 0.795.

3.2 Bilingual Results

It was also attempted to use information from the English language for extrapolating ratings to other languages, specifically Portuguese, Spanish, Italian and German, later leveraging adaptations of the original ANEW dataset into these four separate languages [Redondo et al., 2007, Soares et al., 2012, Montefinese et al., 2014, Schmidtke et al., 2014] in order to evaluate the proposed approach.

Representations were used for the English words in the set of norms from Warriner et al. [2013], specifically for words that do not appear in the ANEW corpora for each target language, as training data for the predictive models.

However, in order to train regression models that can later be used for extrapolating ratings to other languages, there is the need to represent words in the target language in the same embedding space as the training data. Taking inspiration on a previous work by Faruqui and Dyer [2014], canonical correlations analysis was used to project two sets of word embeddings, trained separately for each language, into a common representation space. Forecasting models were trained with basis on the re-projected word embeddings for the English words in the study by Warriner et al. [2013], and they were then applied to the re-projected embeddings for words in the Spanish [Redondo et al., 2007], Portuguese [Soares et al., 2012], Italian [Montefinese et al., 2014] and German [Schmidtke et al., 2014] versions of ANEW. The same forecasting models described in the last subsection were used in this particular set of experiments.

Table 2 presents the results obtained in this second set of experiments. Table 2 also shows the number of words used in the training and testing of each model, as well as the number of words in the seed set

	Number of Words			**k-NN**		
	Training	Testing	Seeds	Valence	Arousal	Dominance
ES	12783	1030	18822	0.613	0.429	0.610
PT	12783	1009	13197	0.565	0.420	0.443
IT	12713	1111	20070	0.586	0.411	0.500
DE	12813	981	9555	0.470	0.332	0.393

	Random Forests			**Kernel Ridge**		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
ES	0.641	0.434	0.596	**0.674**	**0.445**	**0.642**
PT	0.583	**0.447**	0.483	**0.659**	0.391	**0.505**
IT	0.619	**0.454**	0.520	**0.637**	0.453	**0.527**
DE	**0.570**	**0.466**	**0.414**	0.567	0.411	0.402

Table 2: Pearson correlations obtained when predicting the ratings in four different adaptations of the ANEW lexicon, namely for the Spanish, Portuguese, Italian and German languages. The corresponding p-values were always lower than 0.001.

of translations for CCA. The obtained results show that relatively high correlations can be achieved for all four languages, although they are inferior to the results obtained for the monolingual setting.

3.3 Predicting Affective Norms for Short Texts

Despite much sentiment analysis research, few previous studies are directly related to predicting responses to text items in a $[1-9]$ scale according to dimensions of valence, arousal and dominance. The following experiment is based on the idea that emotional norms for English words, such as those made available by Warriner et al. [2013], can be used as training data for predictive models capable of estimating valence, arousal and dominance for previously unseen (sequences of) words.

Following the success of word embedding techniques such as those from word2vec, researchers have tried to extend these models to go beyond the word level, specifically aiming to achieve phrase-level or sentence-level representations. Authors like Le and Mikolov [2014] have proposed more sophisticated approaches, consisting of unsupervised frameworks, similar to those from word2vec, that learn representations for variable-length pieces of text.

In this third experiment, predictive models were trained, using the same aforementioned algorithms (i.e., k-NN, Kernel Ridge and Random Forests) and the lexical norms made available by Warriner et al. [2013], for the prediction of emotional ratings for entire documents. *Paragraph vectors* were trained for the document collections used in our experiments. The implementation from the GenSim package[1] was used for training the *paragraph vectors*, using the default parameters from this library. Prior to model training, the *paragraph vectors* were initialized according to a weighted average of the embeddings for the words occurring in the documents, using TF-IDF scores as the word weights.

To see if the norms made available by Warriner et al. [2013], could be used to estimate valence, arousal and dominance for entire documents, were trained predictive models using the entire dataset of 13,915 English words from Warriner et al. [2013]. Then, the approach was first evaluated through the Affective Norms for English Text (ANET) dataset, which provides a set of normative emotional ratings for a total of 120 brief texts in the English language [Bradley and Lang, 2007]. The results, although inferior to those obtained for predicting the ratings of individual words, attest to the high effectiveness of this method. In fact, the results obtained by this method are significantly superior to those from previous comparable studies [Murphy, 2014]. It should nonetheless be noted that the correlations between the ratings for male and female subjects, in the original ANET study, are of approximately 0.98 for valence, and of 0.96 for both arousal and dominance. This suggests that there is still room for significant improvements.

Later on, to further evaluate this method, was used a previously available dataset, consisting of 20 forum posts rated according to valence and arousal [Paltoglou et al., 2013]. Table 3 presents the obtained results, where we can again see that the proposed methodology is able to rate the texts from this last dataset with a reasonably high accuracy. The method based on embeddings outperforms the results reported by Paltoglou et al. [2013] for the dimension of arousal, although not in terms of valence.

[1] `http://radimrehurek.com/gensim/`

	Forum Posts			
	Valence		Arousal	
	Pearson	MAE	Pearson	MAE
Rnd. Forest	0.753	1.338	0.654	0.667
***k*-NN**	0.627	1.435	*0.193*	0.851
Kernel Ridge	0.791	1.720	0.785	0.377

Table 3: Results obtained when predicting ratings for texts from forum posts. The p-values for the Pearson correlations were lower than 0.001, except for the case shown in italics.

	ANET					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.594	2.201	0.419	2.476	0.583	1.540
***k*-NN**	0.647	2.090	0.569	2.351	0.627	1.499
Kernel Ridge	0.732	1.809	0.649	2.067	0.680	1.392

	EmoTales					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.275	0.999	0.134	1.472	*0.045*	0.876
***k*-NN**	0.330	1.008	0.177	1.424	0.100	0.872
Kernel Ridge	0.347	1.087	0.217	1.418	0.112	0.880

Table 4: Results when predicting ratings for the texts in the ANET [Bradley and Lang, 2007] and EmoTales [Francisco et al., 2012] datasets. The p-values for Pearson's correlations were always lower than 0.0001, except for the case shown in italics.

Finally, in a third set of experiments, was used the *EmoTales* dataset composed of a total of 1168 sentences belonging to 18 folk tales. Table 4 presents the obtained results when predicting the ratings of the sentences that compose the *EmoTales* dataset [Francisco et al., 2012]. The results were significantly lower than the ones obtained on the previous experiments presented in this article. These inferior results, may be assigned to the fact that in this dataset the meaning of each individual sentence, and thereafter its emotional rating is not independent from the other sentences that compose the tale.

4 Affect and Emotions over Twitter Messages

This section describes two different applications that leverages on the prediction of emotional scores (i.e., Valence, Arousal and Dominance) for Tweets. The first application, tries to predict the well-being of populations across continental USA (excluding Alaska), making use of geo-referenced tweets. The second application, consists in seeing if there is correlation between the emotional scores of a given tweet and the metrics associated to the diffusion of that same tweet across the social network.

Both applications rely on the usage of a dataset of Tweets, and in the prediction of emotional scores in terms of Valence, Arousal and Dominance to the textual content of these same tweets. An already existent dataset of 325.333.833 tweets posted by 19.558.917 different users was leveraged on these applications. The tweets in this dataset were collected from the Twitter streaming API service, during the year of 2012. One of the divergence points between the two applications, is related to the set of tweets used in the experiments. Both applications use only the subset of tweets which are written in English (i.e., which have the tag *en* associated to the language of the author). But while the second application makes usage of all the tweets in the aforementioned dataset, the first application uses only those messages which already have an associated pair of coordinates, and whose coordinates lie within the continental USA.

In order to gather all the pre-requisites to predict the emotional scores of the tweets, paragraph vectors were trained for the set of tweets considered by the two applications. The implementation used to train these paragraph vectors was also the one from GenSim package. Having the paragraph vectors associated to the aforementioned set of tweets, and following the methodology presented in the last section where predictive models of the emotional scores for unseen documents were built, was trained a Kernel Ridge regression model using the lexicon from Warriner et al. [2013] and the embeddings from the previously trained paragraph vectors model. This way, the emotional score of each tweet considered for this experiment, was predicted, and able to be used in the two applications

4.1 Predicting Well-Being with Twitter

The first application proposes on leveraging geo-referenced social media data extracted from Twitter within the year of 2012, together with features based on embeddings of tweets and on emotional scores predicted to these same tweets, to estimate well-being of populations from specific geographic points, across continental USA.

As an extension to the approach followed by Loff et al. [2015], the general approach consisted in following a leave-one-out cross validation methodology to train and evaluate a linear regression model that would predict for each state (in continental USA, excluding Alaska) the well-being overall score in the individual state-level reports of the Gallup-Healthways well-being composite index, relative to the year of 2012.

Leveraging both on the paragraph vectors, and on the scores estimated for each one of the considered tweets issued from a given state it was able to produce features that would later be used in the well-being predictive models. Leveraging on the aforementioned features for each one of the states in the continental USA (excluding Alaska), and using the standard scores for each state in the Gallup-Healthways composite well-being index, a linear regression model with Elastic Net regularization was learned, with the aim of predicting well-being scores over the aforementioned states.

The predictions made by the aforementioned models and the ground-truth were then compared by leveraging on the following metrics and following a leave-one-out cross validation methodology: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson correlation coefficient ρ and Kendall correlation coefficient τ. The obtained results (i.e., a MAE of 0.839, a RMSE of 1.163, a ρ of 0.772, and finally a τ of 0.590.), are consistently superior (except in the case of RMSE) than those reported by Loff et al. [2015], i.e, a MAE of 0.92, a RMSE of 1.22, a ρ of 0.74, and finally a τ of 0.58. These results suggest that the method of inferring well-being of populations based on embeddings of geo-referenced tweets is robust.

4.2 Correlating Information Propagation with Affective Ratings

Seeing if the emotional scores of messages (i.e., in terms of valence, arousal, and dominance) affects the way these same messages spread (e.g., how far they get) in social networks like Twitter is the goal of the second application.

The reconstruction of the diffusion processes of a given message imposes particular challenges, since the relations of causality between retweeted messages, needed to establish how the messages spread, are not explicitly expressed in a dataset of messages extracted from Twitter. In Twitter, a retweeted message does not contain in its metadata the id of the message from where it has been retweeted from. This leads to a set of possible messages from which the author may have retweeted. Therefore, to reconstruct the diffusion process, an algorithm based on a set of heuristics must be employed, corresponding to a propagation model. This set of heuristics ranges from the geographical proximity between the author and other users, to their historical interactions (for more details on this algorithm please refer to the full-version of the dissertation).

Regarding the methodology followed to pursue the aforementioned goal, it began by gathering the predictions of the emotional scores relative to the tweets (from the aforementioned dataset) that happened to generate retweets. Then, leveraging on a algorithm which tries to reconstruct the information propagation process (i.e., the chain of retweets) generated by a given tweet, some fluxes of retweets are reconstructed, and on top of these fluxes, diffusion metrics (e.g., depth, width, geographic range, etc) are calculated. Finally, having the emotional scores associated to a tweet that happened to generate a information propagation process, and the metrics associated to this same process, for the set of all reconstructed processes, the relations between each one of the considered emotional dimensions (i.e., valence, arousal, and dominance) and each one of the considered diffusion metrics, were plotted.

By visually analyzing the plots, was not possible to see any obvious correlation between any emotion dimension and any diffusion metric. It is, nevertheless possible to infer that the emotional dimensions follow, as expected, a normal distribution, and that most diffusion metrics follow a exponential distribution, in which there high occurrences for low values, and very low occurrences for high values.

To conclude, we cannot draw strong conclusions, since apparently there is no strong relationship between these two factors in analysis. It is worth to point out that one of the main limitations in the methodology followed to study these relationships, was the fact that the method for reconstructing the path of diffusion of a given tweet across Twitter was not able to be tested. One way of testing it would be to apply this same method to datasets of messages from other social networks where these diffusion paths are explicit. Unfortunately, access to that datasets was impossible.

5 Conclusion

The most valuable contribution and the focus in this work was the building and evaluation of a method for predicting emotional scores, in terms of valence, arousal and dominance for words and short texts based on neural embeddings. The method consisted in building regression models that based on neural embeddings of words or short texts would predict emotional scores in the aforementioned emotional dimensions.

This method produced state-of-the-art results when predicting scores for words in English and significant correlations when predicting scores for words in other languages. The results obtained when evaluating the prediction of scores for short texts, were under the results obtained when predicting scores for words, besides having significant correlation when predicting scores for some datasets.

Finally, this work also contributed with two different applications which leverage on the method for predicting emotional scores for short texts. The first one, based on the embeddings and scores predicted for each one of tweets issued from the continental USA states, predicts the well-being score for each one of these states. These predictions were found to be more robust than previous comparable studies.

The other application, which consisted in seeing if there was a correlation between the emotional score predicted to a given tweet, and the way this same tweet spreads in Twitter according to some diffusion metrics, produced inconclusive results, and no evident relationships between the emotional dimensions and diffusion metrics were found.

Despite the interesting results, there are also several possible paths for improvement such as: Experimenting with alternative word embeddings procedures [Liu et al., 2015, Pennington et al., 2014]; Testing the prediction of scores for short texts with datasets of bigger size; Predicting emotional scores for short texts to other languages than English; And finally, transforming embeddings from different languages to the same vector space leveraging on methods capable of revealing non-linear relations [Andrew et al., 2013, Lopez-Paz et al., 2014].

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, 2013.

M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

Y. Bestgen and N. Vincze. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, 44(4), 2012.

David M. Blei, Andrew Y. Ng, and Michael I.2 Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1), 2003.

M. M. Bradley and P. J. Lang. Affective norms for english words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, 1999.

M. M. Bradley and P. J. Lang. Affective norms for english text (ANET): Affective ratings of text and instruction manual. Technical Report D-1, 2007.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS ONE*, 6(2), 2011.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2014.

Virginia Francisco, Raquel Hervás, Federico Peinado, and Pablo Gervás. Emotales: creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, 46(3), 2012.

Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan. Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In *Computing Attitude and Affect in Text: Theory and Applications*. 2006.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*, 2014.

Jiwei Li, Xun Wang, and Eduard Hovy. What a nasty day: Exploring mood-weather relationship from Twitter. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 2014.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. In *Proccedings of the AAAI Conference on Artificial Intelligence*, 2015.

João Loff, Manuel Reis, and Bruno Martins. Predicting well-being with geo-referenced data collected from social media platforms. In *Proceedings of the ACM/SIGAPP Symposium On Applied Computing*, 2015.

David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. In *Proceedings of the International Conference on Machine Learning*, 2014.

Paweł Manderaa, Emmanuel Keuleersa, and Marc Brysbaerta. How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, 2015.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at the International Conference on Learning Representations*, 2013.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. The adaptation of the affective norms for english words (ANEW) for italian. *Behavior Research Methods*, 46(3), 2014.

Sean Murphy. Validation of the affective norms for english words (ANEW) as weighted computerized language measures of arousal, valence and dominance. Technical report, Pacella Research Center of the New York Psychoanalytic Society and Institute, 2014.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1), 2013.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(1), 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Conference on Empiricial Methods in Natural Language Processing*, 2014.

Guillaume Pitel and Gregory Grefenstette. Semi-automatic building method for a multidimensional affect dictionary for a new language. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

Gabriel Recchiaa and Max M. Louwerse. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68 (12), 2014.

J. Redondo, I. Fraga, I. Padrón, and M. Comesana. The spanish adaptation of ANEW (affective norms for english words). *Behavior Research Methods*, 39(3), 2007.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. ANGST: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior Research Methods*, 46(4), 2014.

A. P. Soares, M. Comesaña, A. P. Pinheiro, A. Simões, and C. S. Frade. The adaptation of the affective norms for english words (ANEW) for european portuguese. *Behavior Research Methods*, 44(1), 2012.

P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 2003.

A.B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 2013.

Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, and Hsuan-Shou Chu. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, 41(0), 2013.

Finn Årup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, 2011.